National Cancer Institute

caBIG™ cancer Biomedical Informatics Grid™

Introduction to the cancer Biomedical Informatics Grid (caBIG™)

George A. Komatsoulis, Ph.D.

Director, Quality Assurance and Compliance

National Cancer Institute Center for Bioinformatics (NCICB)

National Cancer Institute

caBIG™ cancer Biomedical Informatics Grid™

What is caBIG™?

- The caBIG™ program is an initiative of the National Cancer Institute (NCI) designed to create a voluntary network of cancer researchers that create and deploy interoperable data and analytical services, with the goal of speeding the delivery of innovative approaches to cancer diagnosis and treatment.

- The caBIG™ program involves a mixture of technology (including Grid computing and advanced semantics) and community building to achieve its program goals.

- More than 800 people from 80 institutions.

National Cancer Institute

caBIG™ cancer Biomedical Informatics Grid™

Core Principles of caBIG™

- Open Source/Open Access
  - Software and data developed under the auspices of the caBIG™ program are released under a non-viral open source license that allows free access to all
  - But commercial products can become caBIG™ compatible and connect to the caBIG™ Grid without losing IP or becoming Open Source
- Open Development
  - caBIG™ development activities and the artifacts of those activities (requirements documents, activity reports, source code, etc.) are available in a publicly accessible repository
- Federation
  - Data systems in caBIG™ are designed and implemented by members of the cancer research community. System designers are expected to exercise their scientific creativity without being unduly restrained by the NCI

National Cancer Institute

caBIG™ cancer Biomedical Informatics Grid™

Courtesy: Charlie Mead

Interoperability

ability of a system to access and use the parts or equipment of another system

Syntactic interoperability

Semantic interoperability

National Cancer Institute

caBIG™ cancer Biomedical Informatics Grid™

Some data on the Grid

```
<Agent>
    <name>Taxol</name>
    <nSCNumber>007</nSCNumber>
</Agent>
```

National Cancer Institute

caBIG™ cancer Biomedical Informatics Grid™

Attribute	Value	NCI Metadata	CIA Metadata
Agent		A chemical compound administered to a human being to treat an existing disease or condition, or prevent the onset of a disease or condition	A sworn intelligence agent; a spy
nSCNumber	007	Identifier given to a chemical compound by the US Food and Drug Administration (FDA) Nomenclature Standards Committee (NSC)	Identifier given to an intelligence agent by the National Security Council
Name	Taxol	Name of a chemical compound given by the NCI Cancer Therapeutics Evaluation Program (CTEP)	Code name given to intelligence agents by the Central Intelligence Agency (CIA)

National Cancer Institute

caBIG™ cancer Biomedical Informatics Grid™

caBIG™ Compatibility Guidelines

- The caBIG™ compatibility guidelines are designed to insure that systems designed in a Federated environment are still interoperable on the caBIG™ Grid, both syntactically and semantically
- Since achieving interoperability is a process, caBIG™ recognizes four levels of compatibility, starting from Legacy (not interoperable) through Bronze, Silver and Gold (fully interoperable)
- caBIG™ compatibility is all about interfaces rather than the scientific content of the system
- The analogy is to a city
  - In cities architects are free to design buildings that perform myriad functions and that take many distinct forms
  - Nevertheless, all buildings in the city are required to conform to certain specifications in order to receive electricity, water, steam, mail, etc.

National Cancer Institute

caBIG™ cancer Biomedical Informatics Grid™

Maturity Model	Legacy	Bronze	Silver	Gold
Interface Integration	*- No Programming interfaces to the system are available. Only local data files in a custom format can be read* *- Some ad hoc data transfer mechanism such as FTP*	*- Provide baseline* programmatic access to data. Data can be read from remote electronic sources or from commonly used file formats Data can be pushed out to from applications to other external data sources*	*- Well-described API's that provide access to data objects.* *- System architecture separated into tiers and interoperable components* *- Data read in from standards-based electronic sources that support standard or commonly used interchange formats* *- Documented component description of the underlying data structures that are accessible* *- Standard messaging systems where appropriate*	*- All features of Silver, plus:* *- Interoperable with data grid architecture to be defined by caBIG* *- Fully componentized provide access to individual resources in the form of grid services*
Vocabularies / Terminologies & Ontologies	*- Free text used throughout for data collection*	*- Use of publicly accessible standardized controlled vocabularies as well as local terminologies*	*- Standard terminologies approved by public standards bodies or the caBIG Vocabulary/CDE Workspace are used for all relevant data collection fields.*	*- All features of Silver, plus:* *- Fully compliant with caBIG recommended standards for vocabulary terminology services and content sources*
Data Elements	*- No Structured metadata is recorded*	*- Some type of metadata describing the information in the system is used for data collection and external reporting. Metadata is retrieved from external repository shared by multiple applications.* *- Common Data Elements should be built using controlled terminology*	*- Use common standard electronic representation for CDE's such as ISO 11179 or comparable standard* *- CDEs are harmonized and re-used from across the Domain Workspace* *- Common Data Elements are built using standard controlled terminologies approved by public standards bodies or the caBIG Vocabulary/CDE Workspace*	*- All features of Silver, plus:* *- Programmatic access to all metadata, including data class descriptions, site and source information, and any other caBIG-defined metadata requirements and use information models* *- Use the caBIG standard or electronic representation of metadata and Common Data Elements*
Information Models	*- No particular information model is used to represent data*	*- Some type of diagrammatic model describing the data relationship is available in electronic format*	*- Information models defined in a standard modeling language such as UML*	*- All features of Silver, plus:* *- Information models are harmonized with other s across the caBIG Domain Workspace*

National Cancer Institute

caBIG™ cancer Biomedical Informatics Grid™

Enabling Technology

- The NCI provides freely available enabling technology for caBIG™ compatibility
- These technologies are distributed under a ‘non-viral’ open source license.
- caCORE
  - Enterprise Vocabulary Services (EVS)
  - Cancer Data Standards Repository (caDSR)
- caCORE Software Development Kit
  - When complete process is followed, the outcome is a caBIG ‘Silver’ compliant data system.

National Cancer Institute

caBIG™ cancer Biomedical Informatics Grid™

Grid Technology in caBIG™

What is a 'Grid'

- "A Grid is a system that coordinates resources that are not subject to centralized control using standard, open, general-purpose protocols and interfaces to deliver nontrivial qualities of service." - Ian Foster **Grid Today**, July 20, 2002

- Grid Technology supplies two useful components to a network of computers:
  - Advertising: Inform the network about the capabilities of new systems
  - Discovery: Allow users to find resources that meet their needs.

- The caGrid project is the 'Grid in caBIG™'; the actual infrastructure that data and analytical services will use to interoperate.

- The current caGrid is version 0.5; construction of caGrid 1.0 is underway.

- The combination of data and analytical service nodes in caBIG™ produced a design that utilizes a variety of standard Grid technologies including the Globus Toolkit and OGSA-DAI, DQP, GRAM, etc.

National Cancer Institute

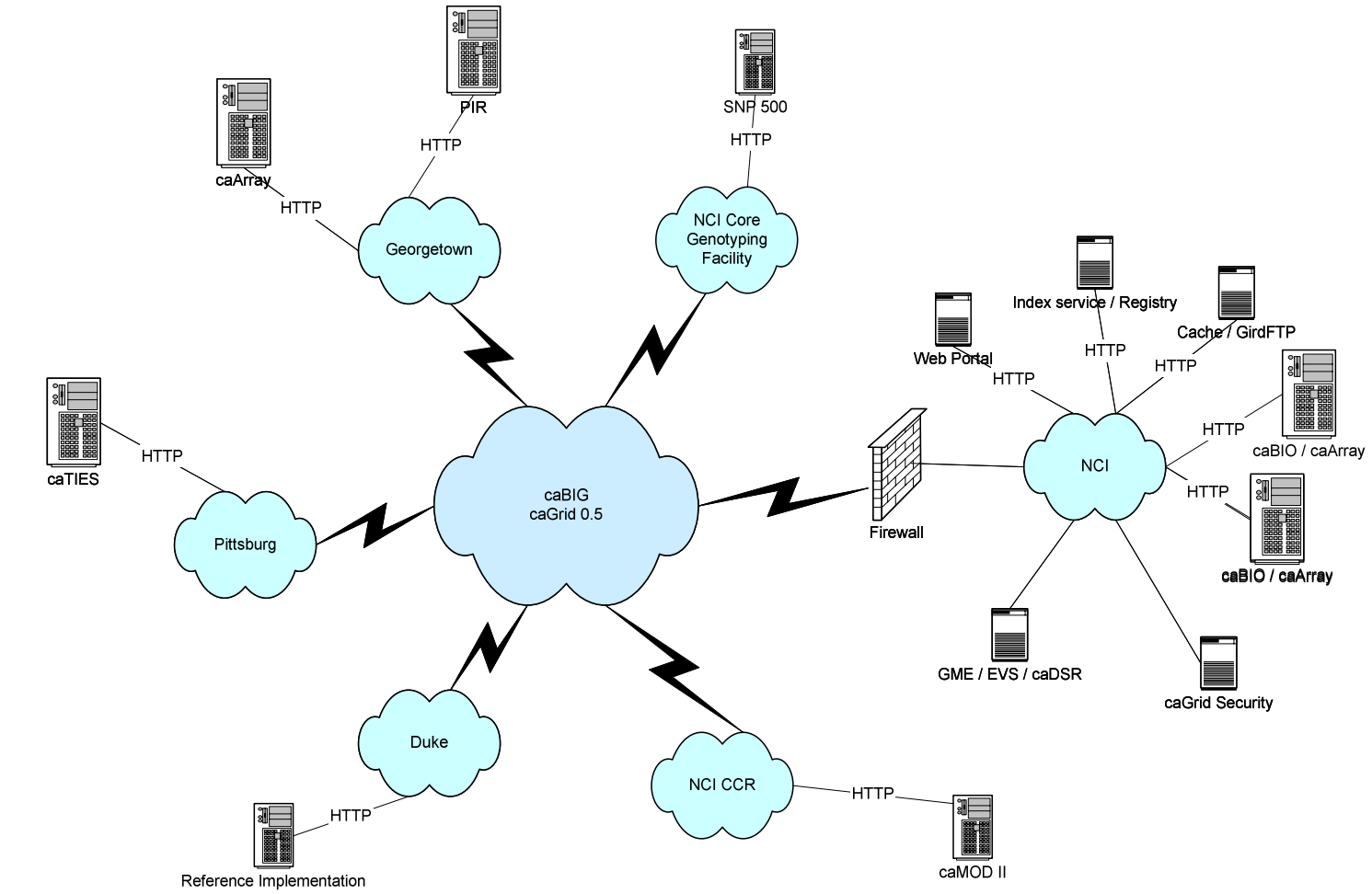
caBIG™ cancer Biomedical Informatics Grid™

Current caGrid 0.5 Nodes

- **Data Services:**
  - caBIO: Gene-centric Bioinformatics Objects (NCICB-Rockville)
  - caArray: Microarray repository (NCICB-Rockville)
  - caArray: Microarray repository (Lombardi Cancer Center-Georgetown)
  - caTIES: Text Information Extraction System for pathology reports (UPMC-Pittsburgh)
  - gridPIR: Protein Information Resource (Lombardi Cancer Center-Georgetown)
  - SNP500: Polymorphism database with population frequencies (NCI Core Genotyping Facility, Gaithersburg, MD)
  - caMOD II: cancer Model Organism Database. (NCI Division of Cancer Research, Rockville, MD)
- **Analytical Services:**
  - RProteomics: Statistical routines for proteomics (Duke-Durham NC)

National Cancer Institute

caBIG™ cancer Biomedical Informatics Grid™



National Cancer Institute

caBIG™ cancer Biomedical Informatics Grid™

Technology Available from caBIG™

- Proteomics Tools:
  - ProtLIMS: Proteomics Laboratory Information Management System (Fox Chase Cancer Center)
  - Q5: Probabilistic Disease Classification of Expression-Dependent Proteomic Data from Mass Spectrometry of Human Serum (Dartmouth)
  - RProteomics: R based statistical tools for Proteomics (Duke Comprehensive Cancer Center)
  - gridPIR: Grid Enablement of the Protein Information Resource (Georgetown/Lombardi Cancer Center)
  - SEED: Genome Analysis framework focused on proteins (University of Chicago)

National Cancer Institute

caBIG™ cancer Biomedical Informatics Grid™

Technology available from caBIG™

- Pathway Tools:
  - cPath: Pathway database that emits information in bioPAX format (Memorial Sloan-Kettering)
  - Reactome: Pathway Data Service (Cold Spring Harbor)
  - QPACA: Quantitative Pathway Analysis in Cancer (University of California, San Francisco)
  - caBIO: Gene centric bioinformatics data (NCICB)

National Cancer Institute

caBIG™ cancer Biomedical Informatics Grid™

Technology available from caBIG™

- Microarray Tools:
  - caArray: MAGE-OM compliant microarray repository (NCICB)
  - Distance Weighted Discrimination: Matlab software for performing systematic bias adjustment in microarray data using Distance Weighted Discrimination (DWD) (UNC-Lineberger Cancer Center)
  - Magellan: A web based system for the analysis of heterogeneous data and annotations (UCSF)
  - Function Express: An automated microarray annotation system (Washington University Siteman Cancer Center)
  - Spot, Sproc: A program for microarray image quantitation (University of California, San Francisco)
  - VISDA: Visual and Statistical Data Analyzer. Provides advanced clustering and other statistical tools for microarray experiments
  - SNP500: Database of polymorphisms with population frequencies (NCI Core Genotyping Facility)

National Cancer Institute

caBIG™ cancer Biomedical Informatics Grid™

For more information

- caBIG™ Web Site
  - http://cabig.nci.nih.gov
- NCICB Web Site
  - http://ncicb.nci.nih.gov
- caCORE
  - http://ncicb.nci.nih.gov/NCICB/infrastructure/cacore_overview
- caCORE SDK
  - http://ncicb.nci.nih.gov/NCICB/infrastructure/cacoresdk
- caBIG™ Compatibility Guidelines
  - https://cabig.nci.nih.gov/guidelines_documentation/caBIGCompatGuideRev2_final.pdf
- caGrid Browser (try out a Grid enabled system)
  - http://cagrid-browser.nci.nih.gov/cagrid-browser/